# Information Theory

***Michele Amoretti***

*High Performance Computing 2022/2023*

# Information theory

Information theory is
- a mathematical theory related to the symbolic aspect of information
- a quantitative approach to the information concept

Information theory answers to the following questions:

- How to store and transmit information in a compact fashion? (**compression**)

- What is the maximum amount of information that can be transmitted on a channel? (**capacity**)

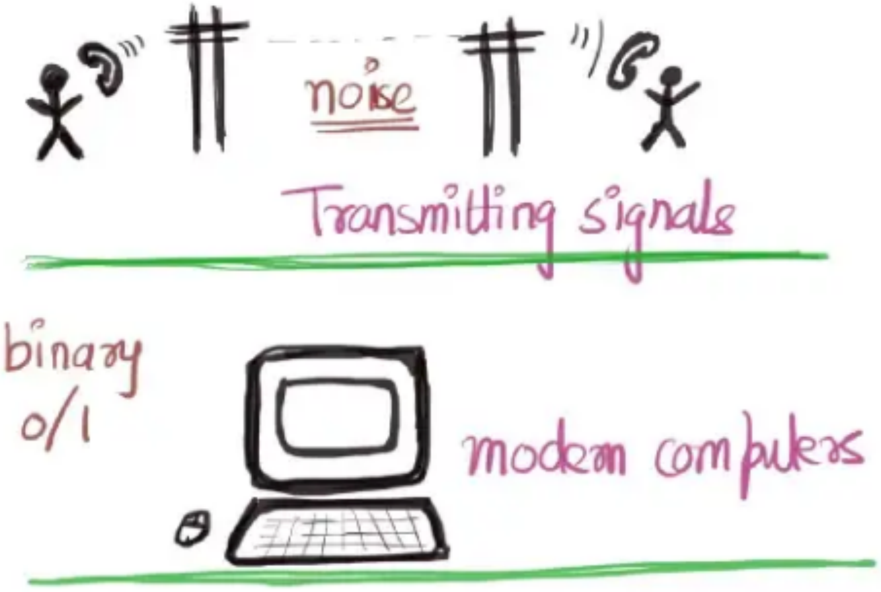
- How can we **protect** our information
  - from corruption or transmission errors?
  - from unauthorized readers?

# Information theory

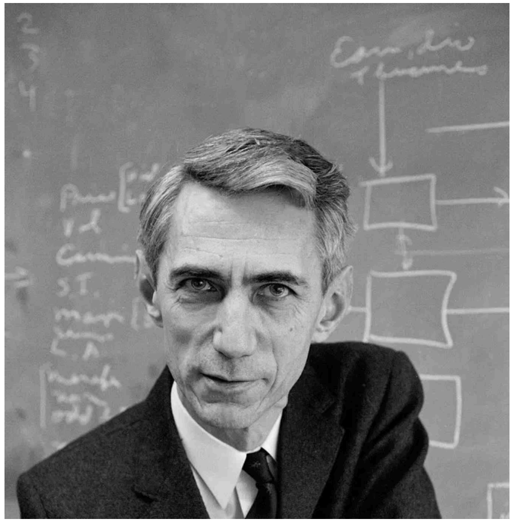
The term **information**, in the context of information theory, has a precise meaning that is somewhat different from our "everyday" experience with it.

Perhaps the word "surprise" better captures the notion of information as it applies in the context of information theory.

In a communication, information is exchanged by means of **messages**.



# Information theory



Information theory was introduced as a scientific subject in 1948 by Claude Shannon.

The first great intuition by Shannon was to consider information as something that helps to answer a question.

To a yes/no (true/false) question, it is always possible to answer with a single two-valued quantity called *binary digit*, or ***bit***.

- physical bit
- information bit

# Information bit

The information bit is a measure of how much we learn from the outcome of a random experiment.

Example: *fair coin*

Without flipping the coin, we have no idea what the result of a coin flip will be. Our best guess at the result is to guess randomly.

If someone else learns the result of a random coin flip, we can ask this person the question:

"What was the result?"

We then learn one bit of information.

# More information bits

What about more complex questions?

“Which one of your four vehicles are you going to use?”

When we get the answer, we learn two bits of information.

**N equally probable outcomes --> $\log_2 N$ bits of information**

# Compression

Attention!

Each digit of a **binary string** can carry (potentially) a bit of information. Nevertheless, it does not always happen.

Ex. The message "0101010101" contains little information because it is merely a repetition of the pattern "01".

A **predictable** message is also a **compressible** message.

To compress a message means to eliminate all redundancy, preserving only the meaningful parts, i.e., information.

The quantity of information contained in a message is the size of the smallest representation of that message.

# Redundancy

Sometimes we add some redundancy to messages, in order to detect and correct errors that are introduced during the transmission phase.

ISBN 978-3-16-148410-0
9 783161 484100

Example: *International Standard Book Number (ISBN) code*

Each ISBN code has 13 digits. The last one is a control digit computed from the previous 12 ones, using the following numerical algorithm:

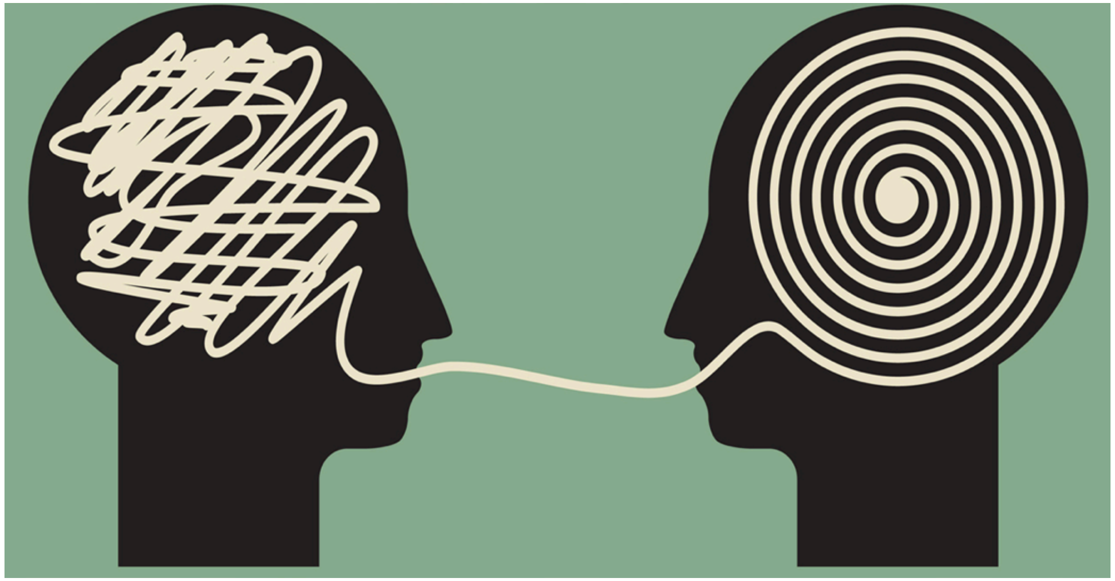
- each digit, from left to right, is multiplied for 1 or 3 alternately
- resulting products are added modulo 10 to obtain a value in [0,9]
- the latter is subtracted from 10, to obtain a result in [1,10]
- if the result is 10, it is substituted with 0

# Redundancy

Redundancy is a **security mechanism**: it guarantees that an information will be received even if the message has been slightly damaged during its transmission.

All languages have similar security mechanisms, made of schemes, structures and sets of rules that make them redundant.

Usually we are not aware of such rules, but our brain learns them automatically, and uses them to control the validity of received messages.



# Capacity of a communication channel

One major success of Shannon's information theory is to have provided a formal definition for redundancy, and to have clearly specified how much information can be transported in a message.

All this is summarized in the **Shannon-Hartley theorem** on the capacity of a communication channel, that was initially conceived to help the engineers of Bell Labs in the evaluation of how many calls could coexist on a phone line, and then used for many other purposes.

$$C = B \log(1 + S/N)$$

$C$ is the channel capacity (b/s), after error correction has been applied.
$B$ is the channel bandwidth (Hz).
$S/N$ is the signal to noise ratio ($S$ e $N$ are powers, expressed in Watt).

**The challenge is to find the best error correction strategy for the given channel, to have the highest bit rate (ideally equal to $C$).**

# Self-information

**Self-information is a measure of the information content of a message $m$.**

Let $p(m) = \mathrm{P}\{m$ out of $\mathcal{M}\}$ be the probability that message $m$ is chosen from all possible choices in the message space $\mathcal{M}$.

$$I(m) = \log(1/p(m)) = -\log p(m)$$

Usually, the log is to the base 2 and self-information is expressed in bits.

Infrequently occurring messages contain more information than more frequently occurring messages.

# Entropy

**Entropy is the average amount of information produced by a stochastic source of data.**

Let $X$ be a discrete random variable with PMF $p(x)$

$$\boldsymbol{H(X) = -\Sigma_x p(x) \log p(x)}$$

Usually, the log is to the base 2. In this case, entropy is expressed in bits.

Properties:

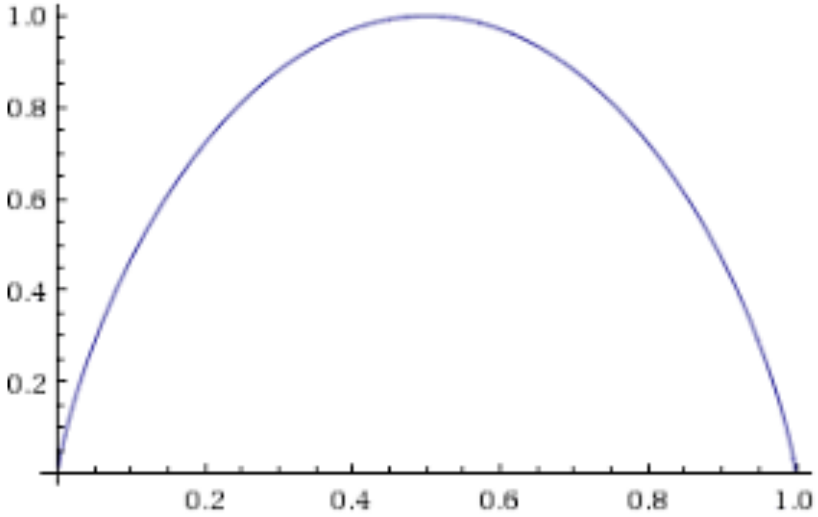
- $H(X) \geq 0$
- $H_m(X) = (\log_m n) H_n(X)$

# Binary entropy function

Let $X$ be a Bernoulli random variable with

$P\{X=1\}=p$

$P\{X=0\}=1-p$



The binary entropy function is

$$H_b(p) = -p\log p -(1-p)\log(1-p)$$

$H_b(p)$ takes a single real number as a parameter, while $H(X)$ takes the PMF of a random variable as a parameter.

$H_b(p)$ is a concave function of the distribution and equals $0$ when $p=0$ or $1$. In this cases the variable is not random and there is no uncertainty. Similarly, the uncertainty is maximum when $p=1/2$, which corresponds to $H_b(p)=1$.

# Joint entropy

$X, Y$ discrete random variables with joint distribution $p(x,y)$

$$\boldsymbol{H(X,Y) = -\Sigma_x \Sigma_y\, p(x,y) \log p(x,y)}$$

# Conditional entropy

It is the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable.

$$H(Y|X) = \Sigma_x\, p(x)\, H(Y|X{=}x) = -\Sigma_x \Sigma_y\, p(x,y) \log p(y|x)$$

Chain rule: $H(X,Y) = H(X) + H(Y|X)$

Remark: $H(X) - H(X|Y) = H(Y) - H(Y|X)$

# Relative entropy

Consider two PMFs $p(x)$ and $q(x)$

$$D(p||q) = \Sigma_x p(x) \log p(x) / q(x)$$

is the relative entropy or Kullback-Leibler distance between the two PMFs. It can be interpreted as the information gain achieved if $q(x)$ is used instead of $p(x)$.

Although it is not a true metric, it has some of the properties of a metric:

- it is always non-negative
- it is zero if and only if $p=q$

but

- it is not symmetric
- it does not satisfy the triangle inequality

# Mutual information

$X,Y$ discrete random variables with joint distribution $p(x,y)$

$$\boldsymbol{I(X;Y) = \Sigma_x \Sigma_y\, p(x,y) \log p(x,y) / p(x)p(y)}$$

**The mutual information is a measure of the dependence between the two random variables.** It is symmetric in $X$ and $Y$ and always non-negative.

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

The formal definition of **channel capacity** for a channel with input $X$ and output $Y$ is $C = \max_{p(x)} I(X;Y)$

# Kolmogorov complexity

**The Kolmogorov complexity measures the information content of a string $x$ by the length of the shortest description $d$ of $x$. Such a length is denoted $C(x)$.**

We think of $d$ as a compressed version of $x$, and the algorithm producing $x$ a decompression procedure.

**A string $x$ is Kolmogorov random if $C(x) \geq |x|$.** Informally, it is the property of $x$ being not longer than any computer program that can produce $x$. In other words, it is the property of $x$ being incompressible.

Kolmogorov complexity has a rich history, with many applications to areas such as computability, machine learning, number theory, and computational complexity.

# References

- C. E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, 1948 http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf
- T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley and Sons, 1991
- M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer, third edition, 2008
- S. Fenner and L. Fortnow, *Compression Complexity*, Technical Report 1702.04779, arXiv.org e-Print archive, 2017